

Equivariant neural networks for inverse problems

Elena Celledoni¹, Matthias J Ehrhardt², Christian Etmann³,
Brynjulf Owren¹, Carola-Bibiane Schönlieb³  and Ferdia
Sherry^{3,*} 

¹ Department of Mathematical Sciences, NTNU, N-7491 Trondheim, Norway

² Institute for Mathematical Innovation, University of Bath, Bath BA2 7JU, United Kingdom

³ Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA, United Kingdom

E-mail: fs436@cam.ac.uk

Received 23 February 2021, revised 8 June 2021

Accepted for publication 30 June 2021

Published 26 July 2021



Abstract

In recent years the use of convolutional layers to encode an inductive bias (translational equivariance) in neural networks has proven to be a very fruitful idea. The successes of this approach have motivated a line of research into incorporating other symmetries into deep learning methods, in the form of group equivariant convolutional neural networks. Much of this work has been focused on roto-translational symmetry of \mathbf{R}^d , but other examples are the scaling symmetry of \mathbf{R}^d and rotational symmetry of the sphere. In this work, we demonstrate that group equivariant convolutional operations can naturally be incorporated into learned reconstruction methods for inverse problems that are motivated by the variational regularisation approach. Indeed, if the regularisation functional is invariant under a group symmetry, the corresponding proximal operator will satisfy an equivariance property with respect to the same group symmetry. As a result of this observation, we design learned iterative methods in which the proximal operators are modelled as group equivariant convolutional neural networks. We use roto-translationally equivariant operations in the proposed methodology and apply it to the problems of low-dose computerised tomography reconstruction and subsampled magnetic resonance imaging reconstruction. The proposed methodology is demonstrated to improve the reconstruction quality of a learned reconstruction method with a little extra computational cost at training time but without any extra cost at test time.

* Author to whom any correspondence should be addressed.



Original content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](https://creativecommons.org/licenses/by/4.0/). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Keywords: image reconstruction, variational regularisation, neural networks, equivariance

(Some figures may appear in colour only in the online journal)

1. Introduction

Deep learning has recently had a large impact on a wide variety of fields; research laboratories have published state-of-the-art results applying deep learning to sundry tasks such as playing Go [1], predicting protein structures [2] and generating natural language [3]. In particular, deep learning methods have also been developed to solve inverse problems, with some examples being [4–7]. In this work we investigate the use of equivariant neural networks for solving inverse imaging problems, i.e. inverse problems where the solution is an image. Convolutional neural networks (CNNs) [8] are a standard tool in deep learning methods for images. By learning convolutional filters, CNNs naturally encode translational symmetries of images: if τ_h is a translation by $h \in \mathbf{R}^d$, and k, f are functions on \mathbf{R}^d , we formally have the following relation (translational equivariance)

$$\tau_h[k * f] = k * [\tau_h f]. \quad (1)$$

This allows learned feature detectors to detect features regardless of their position (though not their orientation or scale) in an image. In many cases it may be desirable for these learned feature detectors to also work when images are transformed under other group transformations, i.e. one may ask that a property such as equation (1) holds for a more general group transformation than the group of translations $\{\tau_h | h \in \mathbf{R}^d\}$. If natural symmetries of the problem are not built into the machine learning method and are not present in the training data, in the worst case, it can result in catastrophic failure as illustrated in figure 1.

To some extent, this problem is circumvented by augmenting the training data through suitable transformations, but it has been shown in classification and segmentation tasks that it is still beneficial to incorporate known symmetries directly into the architecture used, especially if the amount of training data is small [9–11]. Furthermore, training on augmented data is not enough to guarantee that the final model satisfies the desired symmetries. There has recently been a considerable amount of work in this direction, in the form of group equivariant CNNs. Most of the focus has been on roto-translational symmetries of images [9, 11–13], though there is also some work on incorporating scaling symmetries [14, 15] and even on equivariance to arbitrary Lie group symmetries [16].

As mentioned before, we will concern ourselves with solving inverse imaging problems: given measurements y that are related to an underlying ground truth image u through a model

$$y = \mathfrak{N}(A(u)), \quad (2)$$

with A the so-called forward operator and \mathfrak{N} a noise-generating process, the goal is to estimate the image u from the measurements y as well as possible. Typical examples of inverse imaging problems include the problem of recovering an image from its line integrals as in computerised tomography (CT) [17], or recovering an image from subsampled Fourier measurements as in magnetic resonance imaging (MRI) [18, 19]. The solution of an inverse problem is often complicated by the presence of ill-posedness: a problem is said to be well-posed in the sense of Hadamard [20] if it satisfies a set of three conditions (existence of a solution, its uniqueness, and its continuous dependence on the measurements), and ill-posed if any of these conditions fail.

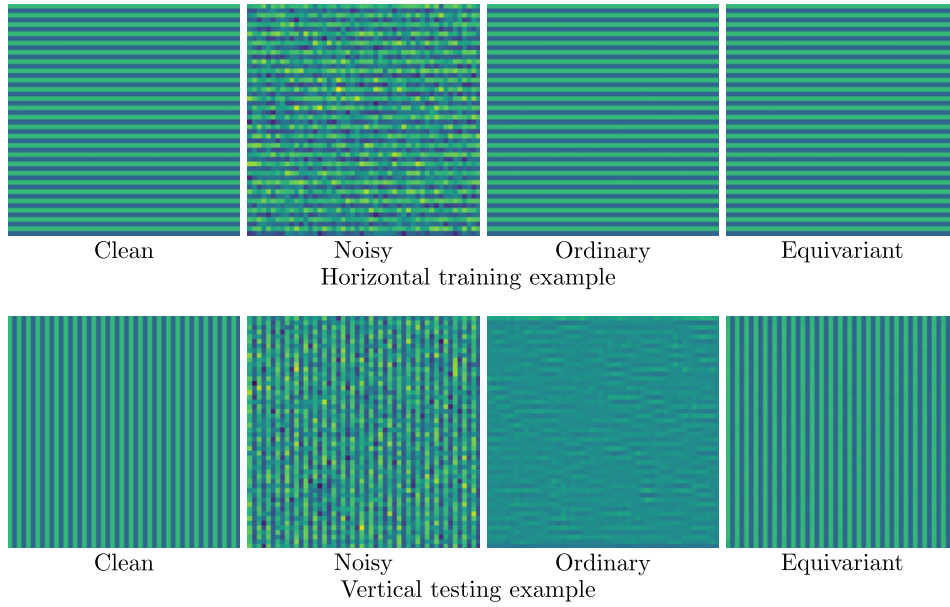


Figure 1. Roto-translationally (‘equivariant’) and just translationally (‘ordinary’) equivariant filters are trained to denoise on a single pair of ground truth and noisy images (‘clean’ and ‘noisy’ in the top row), giving perfect denoising results on the training example. In the bottom row, we see the result of testing the learned filters on a rotated version of the training image; the ordinary filter completely fails at recovering the ground truth, whereas the equivariant filter performs as well as it did on the training image.

It is a natural idea to try to apply equivariant neural networks to solve inverse imaging problems: there is useful knowledge about the relationship between a ground truth image and its measurements in the form of A and the symmetries in both the measurement and image domain (the range and domain of A respectively). Furthermore, training data tends to be considerably less abundant in medical and scientific imaging than in the computer vision and image analysis tasks that are typical of the deep learning revolution, such as ImageNet classification [21]. This suggests that the lower sample complexity of equivariant neural networks (as compared to ordinary CNNs) may be harnessed in this setting with scarce data to learn better reconstruction methods. Finally, end users of the methods, e.g. medical practitioners, are often skeptical of ‘black-box’ methods and guarantees on the behaviour of the method, such as equivariance of the method to certain natural image transformations, may alleviate some of the concerns that they have.

We investigate the use of equivariant neural networks within the framework of learned iterative reconstruction methods [5, 22], which constitute some of the most prototypical deep learning solutions to inverse problems. The designs of these methods are motivated by classical variational regularisation approaches [23, 24], which propose to overcome the ill-posedness of an inverse problem by estimating its solution as

$$\hat{u} = \arg \min_u d(A(u), y) + J(u), \quad (3)$$

with d a measure of discrepancy motivated by our knowledge of the noise-generating process \mathfrak{N} and J is a regularisation functional incorporating prior knowledge of the true solution. Learned

iterative reconstruction methods, also known as unrolled iterative methods, are designed by starting from a problem such as problem (3), choosing an iterative optimisation method to solve it, truncating that method to a finite number of iterations, and finally replacing parts of it (e.g. the proximal operators) by neural networks. We will show that these neural networks can naturally be chosen to be equivariant neural networks, and that doing so gives improved performance over choosing them to be ordinary CNNs. More precisely, our contributions in this work are as follows:

1.1. Our contributions

We show that invariance of a functional to a group symmetry implies that its proximal operator satisfies an equivariance property with respect to that group. This insight can be combined with the unrolled iterative method approach: it makes sense for a regularisation functional to be invariant to roto-translations if there is no prior knowledge on the orientation and position of structures in the images, in which case the corresponding proximal operators are roto-translationally equivariant.

Motivated by these observations, we build learned iterative methods using roto-translationally equivariant building blocks. We show in a supervised learning setting that these methods can outperform comparable methods that only use ordinary convolutions as building blocks, when applied to a low-dose CT reconstruction problem and a subsampled MRI reconstruction problem. This outperformance is manifested in two main ways: the equivariant method is better able to take advantage of small training sets than the ordinary one, and its performance is more robust to transformations that leave images in orientations not seen during training.

2. Notation and background on groups and representations

In this section, we give an overview of the main concepts regarding groups and representations that are required to follow the main text. By a group G , we mean a set equipped with an associative binary operation $\cdot : G \times G \rightarrow G$ (usually the dot is omitted in writing), furthermore containing a neutral element e , such that $e \cdot g = g \cdot e = g$ for all $g \in G$ and a unique inverse g^{-1} for each group element g , such that $g \cdot g^{-1} = g^{-1} \cdot g = e$. Given groups G and H , we say that a map $\phi : G \rightarrow H$ is a group homomorphism if it respects the group structures:

$$\phi(g_1 g_2) = \phi(g_1) \phi(g_2) \quad \text{for any } g_1, g_2 \in G.$$

Groups can be naturally used to describe symmetries of mathematical objects through the concept of group actions. Given a group G and set X , we say that G acts on X if there is a function $T : G \times X \rightarrow X$ (the application of which we stylise as $T_g[x]$ for $g \in G, x \in X$) that obeys the group structure in the sense that

$$T_{g_1} \circ T_{g_2} = T_{g_1 g_2} \quad \text{for any } g_1, g_2 \in G \tag{4}$$

and $T_e = \text{id}$. That is, the group action can be thought of as a group homomorphism from G to the permutation group of X . If there is no ambiguity, the group action may just be written as $T_g[x] = g \cdot x = gx$. An important type of group actions is given by the group representations. If V is a vector space, we will denote by $\text{GL}(V)$ its general linear group, the group of invertible linear maps $V \rightarrow V$, with the group operation given by composition. A representation $\rho : G \rightarrow \text{GL}(V)$ of a group G which acts on V is a group homomorphism, and so corresponds to a linear group action T of G on V : $\rho(g)x = T_g[x]$ for $x \in V$ and $g \in G$. Given a vector space

V , any group G has a representation on V given by $\rho(g) = I$, which is the so-called trivial representation. If V is additionally a Hilbert space, we will call ρ a unitary representation if $\rho(g)$ is a unitary operator for each $g \in G$, i.e. $\|\rho(g)x\| = \|x\|$ for all $x \in V$. Given a finite group $G = \{g_1, \dots, g_n\}$, we can define the so-called regular representation ρ of G on \mathbf{R}^n by

$$\rho(g_i)e_j = e_k,$$

where $\{e_1, \dots, e_n\}$ is a basis of \mathbf{R}^n and k is such that $g_i g_j = g_k$. With this representation, each $\rho(g)$ is a permutation matrix, so ρ is a unitary representation if the basis $\{e_1, \dots, e_n\}$ is orthonormal.

In this work, the groups that we will consider take the form of a group of isometries on \mathbf{R}^d . These groups are represented by a semi-direct product $G = \mathbf{R}^d \rtimes H$, where H is a subgroup of the orthogonal group $O(d)$ of rotations and reflections:

$$O(d) = \{R \in GL(\mathbf{R}^d) | R^T = R^{-1}\}.$$

An important subgroup of $O(d)$ is the special orthogonal group $SO(d) = \{A \in O(d) | \det(A) = 1\}$, which represents the set of pure rotations in $O(d)$. Each element of the semi-direct product G can be identified with a unique pair (t, R) of $t \in \mathbf{R}^d$, the translation component, and $R \in H$, the rotation (and potentially reflection). The semi-direct product can naturally be encoded as a matrix using homogeneous coordinates

$$(t, R) \leftrightarrow \begin{pmatrix} R & t \\ 0 & 1 \end{pmatrix},$$

so that the group product is given by a matrix product. G naturally acts on a point $x \in \mathbf{R}^d$ through $T_{(t,R)}[x] = (t, R)x = Rx + t$.

In the experiments that we consider later in this work, we will consider the case $d = 2$. In this case $SO(2)$ has a simple description:

$$SO(2) = \left\{ \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \mid \theta \in [0, 2\pi) \right\}.$$

We will identify the groups \mathbf{Z}_m of integers modulo m with the subgroup of $SO(2)$ given by

$$\mathbf{Z}_m = \left\{ \begin{pmatrix} \cos(2\pi k/m) & -\sin(2\pi k/m) \\ \sin(2\pi k/m) & \cos(2\pi k/m) \end{pmatrix} \mid k \in \mathbf{Z} \right\}.$$

Given vector spaces V_1, V_2 , we will denote by $\text{Hom}(V_1, V_2)$ the vector space of linear operators $A : V_1 \rightarrow V_2$. We will refer to a number of function spaces: $L^2(\mathbf{R}^d, \mathbf{R}^c)$ denotes the Hilbert space of square integrable functions $f : \mathbf{R}^d \rightarrow \mathbf{R}^c$ (where \mathbf{R}^c carries the Euclidean norm), identified as usual up to equality almost everywhere, and $C_c^\infty(\mathbf{R}^d, \mathbf{R}^c)$ denotes the vector space of infinitely smooth functions $f : \mathbf{R}^d \rightarrow \mathbf{R}^c$ that have compact support.

3. Learnable equivariant maps

The concept of equivariance is well-suited to describing the group symmetries that a function might obey:

Definition 1. Given a general group G , a function $\Phi : \mathcal{X} \rightarrow \mathcal{Y}$ and group actions $T^{\mathcal{X}}, T^{\mathcal{Y}}$ of G on \mathcal{X} and \mathcal{Y} , Φ will be called equivariant if it satisfies

$$\Phi(T_g^{\mathcal{X}}[f]) = T_g^{\mathcal{Y}}[\Phi(f)] \quad (5)$$

for all $f \in \mathcal{X}$ and $g \in G$.

Following the definition of equivariance, we see that equivariant functions have the convenient property that composing them results in an equivariant function, as long as the group actions on the inputs and outputs match in the appropriate way:

Lemma 1. Suppose that G is a group that acts on sets \mathcal{X}, \mathcal{Y} and \mathcal{Z} through $T^{\mathcal{X}}, T^{\mathcal{Y}}$ and $T^{\mathcal{Z}}$. If $\Phi : \mathcal{X} \rightarrow \mathcal{Y}$ and $\Psi : \mathcal{Y} \rightarrow \mathcal{Z}$ are equivariant, then so is $\Psi \circ \Phi : \mathcal{X} \rightarrow \mathcal{Z}$.

Based on this property it is clear that the standard approach to building neural networks (compose linear and nonlinear functions with learnable components in an alternating manner) can be used to build equivariant neural networks as long as linear and nonlinear functions with the desired equivariance can be constructed.

Example 1. Suppose that $\mathcal{X} = L^2(\mathbf{R}^d, \mathbf{R}^{c_{\mathcal{X}}})$ and $\mathcal{Y} = L^2(\mathbf{R}^d, \mathbf{R}^{c_{\mathcal{Y}}})$, with the group $G = \mathbf{R}^d$ acting on \mathcal{X} by $T_h^{\mathcal{X}}[f](x) = f(x - h)$, and in a similar way on \mathcal{Y} by $T^{\mathcal{Y}}$. Ordinary CNNs [8], with convolutional linear layers and pointwise nonlinear functions, are equivariant in this setting.

In this work, we will consider the group $G = \mathbf{R}^d \rtimes H$ for some subgroup H of $O(d)$ (see section 2 for some background), acting on vector-valued functions. To be more specific, we will let $\mathcal{X} = L^2(\mathbf{R}^d, \mathbf{R}^{d_{\mathcal{X}}})$ be the Hilbert space of square-integrable $\mathbf{R}^{d_{\mathcal{X}}}$ -valued functions and assume that $\mathbf{R}^{d_{\mathcal{X}}}$ carries a representation $\pi_{\mathcal{X}} : H \rightarrow GL(\mathbf{R}^{d_{\mathcal{X}}})$. Similarly, we will define $\mathcal{Y} = L^2(\mathbf{R}^d, \mathbf{R}^{d_{\mathcal{Y}}})$ and assume that $\pi_{\mathcal{Y}} : H \rightarrow GL(\mathbf{R}^{d_{\mathcal{Y}}})$ is a representation of H . We define the group actions $T^{\mathcal{X}}$ and $T^{\mathcal{Y}}$ to be the induced representations, $\rho_{\mathcal{X}}$ and $\rho_{\mathcal{Y}}$, of $\pi_{\mathcal{X}}$ and $\pi_{\mathcal{Y}}$ on \mathcal{X} and \mathcal{Y} respectively. In the setting that we are considering, these representations take a particularly simple form. As mentioned in section 2, since we assume that G takes the semi-direct product form $\mathbf{R}^d \rtimes H$, each group element $g \in G$ can be uniquely thought of as a pair $g = (t, R)$ for some $t \in \mathbf{R}^d$ and $R \in H$. With this in mind, the representations $\rho_{\mathcal{X}}$ and $\rho_{\mathcal{Y}}$ can be written as follows for any $f \in \mathcal{Z}, x \in \mathbf{R}^d$ and $t \in \mathbf{R}^d, R \in H$:

$$\rho_{\mathcal{Z}}((t, R))[f](x) = \underbrace{\pi_{\mathcal{Z}}(R)}_{(a)} \underbrace{f((t, R)^{-1}x)}_{(b)} \quad \text{for } \mathcal{Z} = \mathcal{X}, \text{ or } \mathcal{Z} = \mathcal{Y}. \quad (6)$$

These representations have a natural interpretation: to apply a group element (t, R) to a vector-valued function, we must move the vectors, as in part (b) of equation (6), and transform each vector accordingly, as in part (a) of equation (6).

3.1. Equivariant linear operators

It is well-established that equivariant linear operators are strongly connected to the concept of convolutions. Indeed, in a relatively general setting it has been shown that an integral operator is equivariant if and only if it is given by a convolution with an appropriately constrained kernel

[25]. In the setting that we are considering, the more specific result in proposition 1 can be derived, as done in [11, 26] for the case $d = 2$ and [27] for the case $d = 3$.

Proposition 1. *Suppose that $\Phi : \mathcal{X} \rightarrow \mathcal{Y}$ is an operator given by integration against a continuous kernel $K : \mathbf{R}^d \times \mathbf{R}^d \rightarrow \text{Hom}(\mathbf{R}^{d_x}, \mathbf{R}^{d_y})$,*

$$\Phi(f)(x) = \int_{\mathbf{R}^d} K(x, y) f(y) dy.$$

Then the operator Φ is equivariant if and only if it is in fact given by a convolution satisfying an additional constraint: there is a continuous $k : \mathbf{R}^d \rightarrow \text{Hom}(\mathbf{R}^{d_x}, \mathbf{R}^{d_y})$

$$\Phi(f)(x) = \int_{\mathbf{R}^d} k(x - y) f(y) dy,$$

where k satisfies the additional condition

$$k(Rx) = \pi_Y(R)k(x)\pi_X(R^{-1}) \quad \text{for } x \in \mathbf{R}^d, R \in H.$$

The derivation of this result proceeds by writing out the definitions of equivariance and using the invariances of the Lebesgue measure. The equivariance of Φ implies that we must have the following chain of equalities for any $x \in \mathbf{R}^d, f \in \mathcal{X}, t \in \mathbf{R}^d, R \in H$ and $g = (t, R) \in G$:

$$\begin{aligned} \int_{\mathbf{R}^d} \pi_Y(R)K(g^{-1}x, y)f(y) dy &\stackrel{(a)}{=} \pi_Y(R) \int_{\mathbf{R}^d} K(g^{-1}x, y)f(y) dy \\ &= \rho_Y(g)[\Phi(f)](x) \\ &\stackrel{(b)}{=} \Phi(\rho_X(g)[f])(x) \\ &= \int_{\mathbf{R}^d} K(x, y)\rho_X(g[f])(y) dy \\ &= \int_{\mathbf{R}^d} K(x, y)\pi_X(h)f(g^{-1}y) dy \\ &\stackrel{(c)}{=} \int_{\mathbf{R}^d} K(x, gy)\pi_X(h)f(y) dy. \end{aligned}$$

Here the tags above the equality signs correspond to the following justifications:

- (a) Since π_Y is a group representation, $\pi_Y(R)$ is a linear map and commutes with the integral,
- (b) Φ is assumed to be equivariant,
- (c) We make the substitution $y \leftarrow gy$ and note that the Lebesgue measure is invariant to G .

Taking the left-hand side and right-hand side together, we find that

$$\int_{\mathbf{R}^d} (\pi_Y(R)K(g^{-1}x, y) - K(x, gy)\pi_X(h)) f(y) dy = 0,$$

and since this must hold for any $f \in \mathcal{X} = L^2(\mathbf{R}^d, \mathbf{R}^{d_{\mathcal{X}}})$, we conclude by testing on sequences converging to Dirac delta functions that

$$\pi_{\mathcal{Y}}(R)K(g^{-1}x, y) = K(x, gy)\pi_{\mathcal{X}}(R). \quad (7)$$

Specialising by setting R equal to the identity element, we see that

$$K(x - t, y) = K((t, I)^{-1}x, y) = K(x, (t, I)y) = K(x, y + t),$$

or upon substituting $x \leftarrow x + t$, $K(x, y) = K(x + t, y + t)$. Choosing t to be the translation that takes y to 0, we find that

$$K(x, y) = K(x - y, 0) =: k(x - y)$$

defines a convolution kernel $k: \mathbf{R}^d \rightarrow \text{Hom}(\mathbf{R}^{d_{\mathcal{X}}}, \mathbf{R}^{d_{\mathcal{Y}}})$. Now specialising equation (7) by letting $R \in H$ and $x \in \mathbf{R}^d$ be arbitrary and $t, y = 0$, we obtain the condition $\pi_{\mathcal{Y}}(R)k(R^{-1}x) = k(x)\pi_{\mathcal{X}}(R)$, or upon substituting $x \leftarrow Rx$ and rearranging,

$$k(Rx) = \pi_{\mathcal{Y}}(R)k(x)\pi_{\mathcal{X}}(R^{-1}). \quad (8)$$

Conversely, the above reasoning can be reversed to show that the condition in equation (8) (for all $x \in \mathbf{R}^d, R \in H$) is sufficient to guarantee equivariance of Φ .

The condition in equation (8) is a linear constraint that is fully specified before training. Hence, if a basis is computed for the convolution kernels satisfying equation (8), a general equivariant linear operator can be learned by learning its parameters in that basis. Since the choices of H that we consider are all compact groups, any representation of H can be decomposed as a direct sum of irreducible representations of H (theorem 5.2 in [28]). As a result of this, we can give the following procedure to compute a basis for the convolution kernels satisfying the equivariance condition in equation (8) as soon as $\pi_{\mathcal{X}}$ and $\pi_{\mathcal{Y}}$ are specified:

- Decompose $\pi_{\mathcal{X}}$ and $\pi_{\mathcal{Y}}$ as direct sum of irreducible representations; $\pi_{\mathcal{X}} = Q_{\mathcal{X}} \text{diag}(\pi_{\mathcal{X}}^1, \dots, \pi_{\mathcal{X}}^{k_{\mathcal{X}}})Q_{\mathcal{X}}^{-1}$, $\pi_{\mathcal{Y}} = Q_{\mathcal{Y}} \text{diag}(\pi_{\mathcal{Y}}^1, \dots, \pi_{\mathcal{Y}}^{k_{\mathcal{Y}}})Q_{\mathcal{Y}}^{-1}$ (here diag constructs a block diagonal matrix with the diagonal elements given by the arguments supplied to diag).
- For each i, j with $1 \leq i \leq k_{\mathcal{X}}, 1 \leq j \leq k_{\mathcal{Y}}$ find a basis for the convolution kernels $k_{i,j}$ satisfying the equivariance condition

$$k_{i,j}(Rx) = \pi_{\mathcal{Y}}^j(R)k_{i,j}(x)\pi_{\mathcal{X}}^i(R^{-1})$$

with the irreducible representations $\pi_{\mathcal{Y}}^j$ and $\pi_{\mathcal{X}}^i$.

- Given expansions of the $k_{i,j}$, compute the overall equivariant convolution kernel k by

$$k = Q_{\mathcal{Y}} \cdot (k_{i,j})_{1 \leq i \leq k_{\mathcal{X}}, 1 \leq j \leq k_{\mathcal{Y}}} \cdot Q_{\mathcal{X}}^{-1}.$$

This procedure has been described in more detail in [11] and implemented in the corresponding software package for the groups $G = \mathbf{R}^2 \rtimes H$, where H can be any subgroup of $O(2)$.

Since the equivariant convolutions described above are implemented using ordinary convolutions, little extra computational effort required to use them compared to ordinary convolutions: during training, there is just an additional step of computing the basis expansion defining the equivariant convolution kernels (and backpropagating through it). When it is time to test the network, this step can be avoided by computing the basis expansion once and only saving

the resulting convolution kernels, so that it is completely equivalent in terms of computational effort to using an ordinary CNN.

3.2. Equivariant nonlinearities

Although pointwise nonlinearities are translationally equivariant, some more care is needed when designing nonlinearities that satisfy the equivariance condition in equation (5) with our choices of groups. Examining the form of the induced representations in our setting, as given in equation (6), it is evident that for a pointwise nonlinearity $\phi : \mathbf{R} \rightarrow \mathbf{R}$ to be equivariant (in the sense that $\phi(\rho_{\mathcal{X}}(g)[f]) = \rho_{\mathcal{X}}(g)[\phi(f)]$, with ϕ applied pointwise) ϕ must commute with $\pi_{\mathcal{X}}(R)$ for every $R \in H$: with $g = (t, R)$ for $t \in \mathbf{R}^d, R \in H$ we have

$$\begin{aligned}\phi(\pi_{\mathcal{X}}(R)f(g^{-1}x)) &= \phi(\rho_{\mathcal{X}}(g)[f])(x) = \rho_{\mathcal{X}}(g)[\phi(f)](x) \\ &= \pi_{\mathcal{X}}(h)\phi(f(g^{-1}x)).\end{aligned}$$

This can be ensured if $\pi_{\mathcal{X}}$ is the regular representation of H , since in that case each $\pi_{\mathcal{X}}(h)$ is a permutation matrix, giving the following guideline:

Lemma 2. *Suppose that $G = \mathbf{R}^d \rtimes H$ with H a finite subgroup of $O(d)$ and that $\phi : \mathbf{R} \rightarrow \mathbf{R}$ is a given function. If $\pi_{\mathcal{X}}$ is the regular representation of H , then $\Phi : \mathcal{X} \rightarrow \mathcal{X}$ is equivariant, where $\Phi(f)(x) = \phi(f(x))$.*

Another way to ensure that ϕ commutes with $\pi_{\mathcal{X}}$ is by choosing the trivial representation. Although the trivial representation may not be very interesting by itself, this gives rise to another form of nonlinearity called the norm nonlinearity. If $\pi_{\mathcal{X}}$ is a unitary representation, taking the pointwise norm satisfies an equivariance condition: with $g = (t, R)$ for $t \in \mathbf{R}^d, R \in H$

$$\|\rho_{\mathcal{X}}(g)[f](x)\| = \|\pi_{\mathcal{X}}(R)f(g^{-1}x)\| = \|f(g^{-1}x)\|.$$

The right-hand side transforms according to the trivial representation, so by the above comments we deduce that the nonlinearity $f \mapsto \phi(\|f\|)$ satisfies an equivariance condition of the same form. To obtain the norm nonlinearity, which maps features of a given type to features of the same type, we then form the map $\Phi : \mathcal{X} \rightarrow \mathcal{X}, f \mapsto f \cdot \phi(\|f\|)$: with $g = (t, R)$ for $t \in \mathbf{R}^d, R \in H$, we have

$$\begin{aligned}\Phi(\rho_{\mathcal{X}}(g)[f])(x) &= \pi_{\mathcal{X}}(R)f(g^{-1}x) \cdot \phi(\|f(g^{-1}x)\|) \\ &= \pi_{\mathcal{X}}(R)(f(g^{-1}x) \cdot \phi(\|f(g^{-1}x)\|)) \\ &= \pi_{\mathcal{X}}(R)(f \cdot \phi(\|f\|))(g^{-1}x) \\ &= \rho_{\mathcal{X}}(g)[\Phi(f)](x),\end{aligned}$$

where we used that $\phi(\|f(g^{-1}x)\|)$ is a scalar. This shows that the norm nonlinearity Φ is indeed equivariant:

Lemma 3. *Suppose that $\pi_{\mathcal{X}}$ is a unitary representation of H , and that $\phi : \mathbf{R} \rightarrow \mathbf{R}$ is a given function. Then the norm nonlinearity $\Phi : \mathcal{X} \rightarrow \mathcal{X}$ with $\Phi(f)[x] = f(x)\phi(\|f(x)\|)$ is equivariant.*

4. Reconstruction methods motivated by variational regularisation

We consider the inverse problem of estimating an image u from noisy measurements y . We will assume that knowledge of the measurement process is available in the form of the forward operator A , which maps an image to ideal, noiseless measurements, and generally there were

Algorithm 1. Proximal gradient method.

inputs: measurements y , initial estimate u^0
 $u \leftarrow u^0$
for $i \leftarrow 1, \dots$, it do
 $u \leftarrow \text{prox}_{\tau J}(u - \tau^i \nabla E_y(u))$
end for
return u

will be a reasonable idea of the process by which they are corrupted to give rise to the noisy measurements y . A tried and tested approach to solving inverse problems is the variational regularisation approach [23, 29]. In this approach, images are recovered from measurements by minimising a trade-off between the data fit and a penalty function encoding prior knowledge:

$$\hat{u} = \arg \min_u E_y(u) + J(u), \quad (9)$$

with E_y a data discrepancy functional penalising mismatch of the estimated image and the measurements and J the penalty function. Usually E_y will take the form $E_y(u) = d(A(u), y)$, where d is a measure of divergence chosen based on our knowledge of the noise process.

4.1. Equivariance in splitting methods

Generally, problem (9) may be difficult to solve, and a lot of research has been done on methods to solve problems such as these. Iterative methods to solve it are often structured as splitting methods: the objective function is split into terms, and easier subproblems associated with each of these terms are solved in an alternating fashion to yield a solution to problem (9) in the limit. A prototypical example of this is the proximal gradient method (also known as forward-backward splitting) [30, 31], which has become a standard tool for solving linear inverse problems, particularly in the form of the FISTA algorithm [32]. In its basic form, the proximal gradient method performs the procedure described in algorithm 1.

Recall here that the proximal operator [33–35] prox_J is defined as follows:

Definition 2. Suppose that \mathcal{X} is a Hilbert space and that $J : \mathcal{X} \rightarrow \mathbf{R} \cup \{+\infty\}$ is a lower semi-continuous convex proper functional. The proximal operator $\text{prox}_J : \mathcal{X} \rightarrow \mathcal{X}$ is then defined as

$$\text{prox}_J(u) = \arg \min_{u' \in \mathcal{X}} \frac{1}{2} \|u - u'\|^2 + J(u'). \quad (10)$$

Although this definition of proximal operators assumes that the functional J is convex, this assumption is more stringent than is necessary to ensure that an operator defined by equation (10) is well-defined and single-valued. One can point for example to the classes of μ -semi-convex functionals (i.e. the set of J , such that $u \mapsto J(u) + \frac{\mu}{2} \|u\|^2$ is convex) on \mathcal{X} for $0 < \mu < 1$, which include nonconvex functionals. In what follows, we will allow for such more general functionals by just asking that the proximal operator is well-defined and single-valued.

It is often reasonable to ask that the proximal operators $\text{prox}_{\tau J}$ satisfy an equivariance property; if the corresponding regularisation functional J is invariant to a group symmetry, the proximal operator will be equivariant:

Proposition 2. Suppose that \mathcal{X} is a Hilbert space and ρ is a unitary representation of a group G on \mathcal{X} . If a functional $J : \mathcal{X} \rightarrow \mathbf{R} \cup \{+\infty\}$ is invariant, i.e. $J(\rho(g)f) = J(f)$, and has a well-defined single-valued proximal operator $\text{prox}_J : \mathcal{X} \rightarrow \mathcal{X}$, then prox_J is equivariant, in

the sense that

$$\text{prox}_J(\rho(g)f) = \rho(g)\text{prox}_J(f)$$

for all $f \in \mathcal{X}$ and $g \in G$.

Proof. We have the following chain of equalities:

$$\begin{aligned} \text{prox}_J(\rho(g)f) &= \arg \min_h \frac{1}{2} \|\rho(g)f - h\|^2 + J(h) \\ &\stackrel{(a)}{=} \arg \min_h \frac{1}{2} \|\rho(g)(f - \rho(g^{-1})h)\|^2 + J(\rho(g^{-1})h) \\ &\stackrel{(b)}{=} \arg \min_h \frac{1}{2} \|f - \rho(g^{-1})h\|^2 + J(\rho(g^{-1})h) \\ &\stackrel{(c)}{=} \rho(g) \left[\arg \min_h \frac{1}{2} \|f - h\|^2 + J(h) \right] = \rho(g)\text{prox}_J(f). \end{aligned}$$

The three marked steps are justified as follows:

- (a) J is assumed to be invariant w.r.t. ρ ,
- (b) The representation ρ is assumed to be unitary,
- (c) $\rho(g)$ is invertible, and under the substitution $h \leftarrow \rho(g)h$, the minimiser transforms accordingly.

□

Example 2. As a prominent example of a regularisation functional satisfying the conditions of proposition 2, consider the total variation functional [36] on $L^2(\mathbf{R}^d)$

$$\text{TV}(u) = \sup_{\phi \in C_c^\infty(\mathbf{R}^d; \mathbf{R}^d), \|\phi\|_\infty \leq 1} \int_{\mathbf{R}^d} u \operatorname{div} \phi,$$

with the group $G = \text{SE}(d)$ and the scalar field representation $\rho(r)[f](x) = f(r^{-1}x)$. Since the Lebesgue measure is invariant to G and the set of vector fields $\{\phi \in C_c^\infty(\mathbf{R}^d; \mathbf{R}^d) \mid \|\phi\|_\infty \leq 1\}$ is closed under G , TV is invariant w.r.t. ρ . As a result of this, proposition 2 tells us that $\text{prox}_{\tau\text{TV}}$ is equivariant w.r.t. ρ for any $\tau \geq 0$. Note that TV is not unique in satisfying these conditions; by a similar argument it can be shown, for example, that the higher order total generalised variation functionals [37] share the same invariance property (and hence also that their proximal operators are equivariant).

Remark 1. The above example, and all other examples that we consider in this work, are concerned with the case where the image to be recovered is a scalar field. Note, however, that proposition 2 is not limited to this type of field and that there are applications where it is natural to use more complicated representations ρ . A notable example is diffusion tensor MRI [38] in which case the image to be estimated is a diffusion tensor field and ρ should be chosen as the appropriate tensor representation.

4.1.1. Equivariance of the reconstruction operator. It is worth thinking about whether it is sensible to ask that the overall reconstruction method is equivariant, and how this should be interpreted. Thinking of the reconstruction operator as a map from measurements y to images \hat{u} , it is hard to make sense of the statement that it is equivariant, since the measurement space

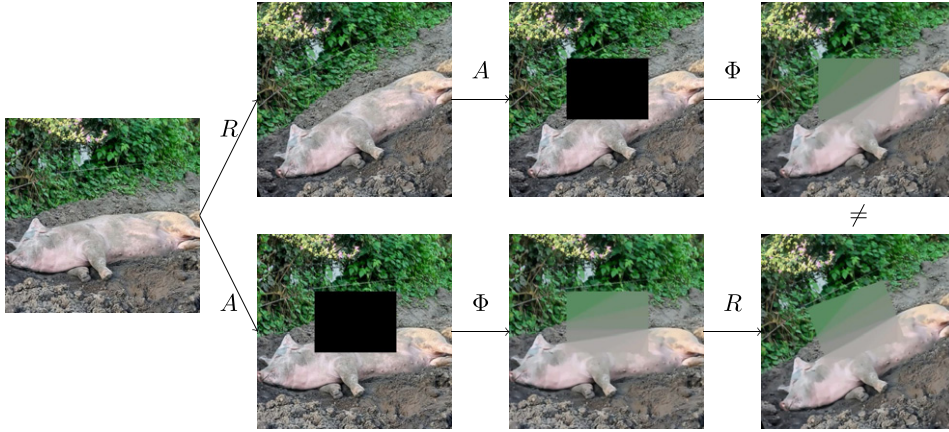


Figure 2. An example demonstrating the non-equivariance of a general variational regularisation approach to image reconstruction, even when the corresponding regularisation functional J (as in problem (9)) is invariant. Here, A represents the application of an inpainting mask, R is an operator rotating the image by 20° and Φ is the solution map to problem (9) with $E_y(u) = \|Au - y\|^2$ and $J(u) = \tau \text{TV}(u)$.

Algorithm 2. Learned proximal gradient method.

inputs: measurements y , initial estimate u^0
 $u \leftarrow u^0, s \leftarrow 0$
for $i \leftarrow 1, \dots, \text{it}$ do
 $(u, s) \leftarrow \widehat{\text{prox}}_i(u, s, \nabla E_y(u))$
end for
return $\Phi(y) := u$

generally does not share the symmetries of the image space (in the case where measurements may be incomplete). If we think instead of the reconstruction method as mapping a true image u to an estimated image \hat{u} through (noiseless) measurements $y = A(u)$, we might ask that a symmetry transformation of u should correspond to the same symmetry transformation of \hat{u} . In the case of reconstruction by a variational regularisation method as in problem (9), this is too much to ask for even if the regularisation functional is invariant, since information in the (incomplete) measurements can appear or disappear under symmetry transformations of the true image. An example of this phenomenon when solving an inpainting problem is shown in figure 2.

4.2. Learned proximal gradient descent

A natural way to use knowledge of the forward model in a neural network approach to image reconstruction is in the form of unrolled iterative methods [5, 22]. Starting from an iterative method to solve problem (9), the method is truncated to a fixed number of iterations and some of the steps in the truncated algorithm are replaced by learnable parts. As noted in the previous section, the proximal gradient method in algorithm 1 can be applied to a variational regularisation problem such as problem (9). Motivated by this and the unrolled iterative method approach, we can study learned proximal gradient descent as in algorithm 2 (where the variable s can be

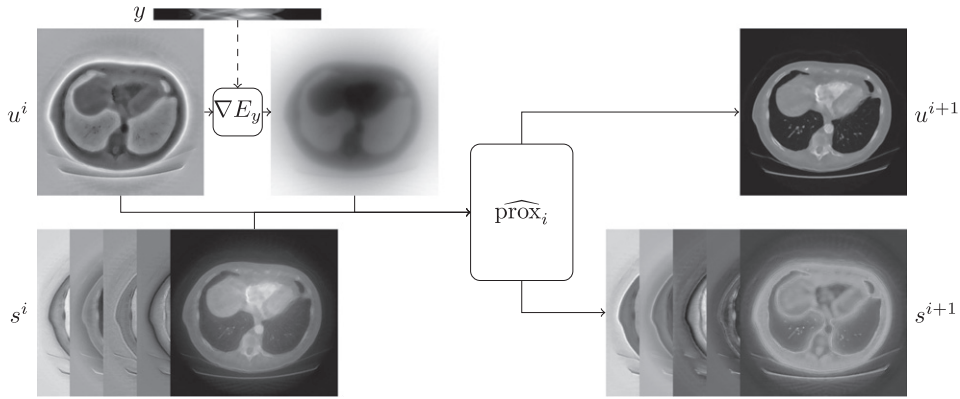


Figure 3. A schematic illustration of a single iteration of the learned proximal gradient method, algorithm 2, for a CT reconstruction problem. The choice of E_y is described in section 5.1.1. Knowledge of the forward model is incorporated into the reconstruction through ∇E_y , which is not an equivariant operator in general. Motivated by proposition 2, we know that $\widehat{\text{prox}}_i$ is naturally modelled as an equivariant operator.

used as a memory state as is common in accelerated versions of the proximal gradient method [32]):

Here $\widehat{\text{prox}}_i$ are neural networks, the architectures of which are chosen to model proximal operators. In this work, we choose $\widehat{\text{prox}}_i$ to be defined as

$$\widehat{\text{prox}}_i = K_{\text{project},i} \circ (\text{id} + \phi \circ K_{\text{intermediate},i}) \circ K_{\text{lift},i}, \quad (11)$$

where each of the $K_{\text{project},i}$, $K_{\text{intermediate},i}$ and $K_{\text{lift},i}$ are learnable affine operators (given by a convolution operation followed by adding a bias term) and ϕ is an appropriate nonlinear function. We can appeal to proposition 2 and model $\widehat{\text{prox}}_i$ as translationally equivariant (we will call the corresponding reconstruction method the ordinary method in what follows) or as roto-translationally equivariant (we will call the corresponding reconstruction method the equivariant method in what follows). Figure 3 gives a schematic illustration of the inputs and outputs of the learned proximal operators.

Recall that we consider groups of the form $G = \mathbf{R}^d \rtimes H$ for subgroups H of $O(d)$ in this work. Since we apply the learned equivariant method to reconstruct scalar-valued images, the input and output types of each $\widehat{\text{prox}}_i$ should correspond to features carrying the trivial representation of H . For the equivariant method, $K_{\text{lift},i}$ are equivariant convolutions from a small number ($2 +$ the number of channels used for the memory state) of input channels with the trivial representation of H to a larger number of intermediate channels with the regular representation of H , if H is a finite group, or various irreducible representations of H , if H is a continuous group. $K_{\text{intermediate},i}$ are chosen as equivariant convolutions mapping the output channels of $K_{\text{lift},i}$ to a set of channels of the same type. Finally, $K_{\text{project},i}$ are chosen as equivariant convolutions that map the output channels of $K_{\text{intermediate},i}$ to a small number ($1 +$ the number of channels used for the memory states) of output channels with the trivial representation of H . For the implementation of the equivariant convolutions, recall the procedure described at the end of section 3.1.

For the ordinary method, $K_{\text{lift},i}$ are ordinary convolutions mapping a small number (equal to that of the equivariant method) of input channels to a larger number of intermediate channels, $K_{\text{intermediate},i}$ are ordinary convolutions mapping the output channels of $K_{\text{lift},i}$ to a set of channels

of the same type, and $K_{\text{project},i}$ are ordinary convolutions mapping the many output channels of $K_{\text{intermediate},i}$ to a small number (equal to that of the equivariant method) of output channels.

Since the implementations of the equivariant convolutions are ultimately based on ordinary convolutions, a natural comparison can be made between the equivariant and ordinary method by matching the widths of the underlying ordinary convolutions. When the methods are compared in this way, they should take comparable computational effort to use and the ordinary method is a superset of the equivariant method in the sense that the parameters of the ordinary method can be chosen to reproduce the action of the equivariant method.

Remark 2. Both in the case of algorithms 1 and 2, we require access to the gradient ∇E_y , where E_y is a data discrepancy functional. In our case, E always takes the form $E_y(u) = d(A(u), y)$ where A is the forward operator and d is a measure of divergence. As a result of this E_y can be differentiated by the chain rule as long as we have access to the gradient of d and can compute vector-Jacobian products of A . If the forward operator A is linear, its vector-Jacobian products are just given by the action of the adjoint of A .

5. Experiments

In this section, we demonstrate that roto-translationally equivariant operations can be incorporated into a learned iterative reconstruction method such as algorithm 2 to obtain higher quality reconstructions than those obtained using comparable reconstruction methods that only use translationally equivariant operations. We consider two different inverse problems: a subsampled MRI problem and a low-dose CT problem. The code that was used to produce the experimental results shown is freely available at https://github.com/fsherry/equivariant_image_recon [54].

5.1. Datasets

5.1.1. LIDC-IDRI dataset. We use a selection of chest CT images of size 512×512 from the LIDC-IDRI dataset [39, 40] for our CT experiments. We use a combination of L^1 norm and the TV functional as a simple way to screen out low-quality images. The details of this procedure can be found in the code repository associated with this work. The set is split into 5000 images that can be used for training, 200 images that can be used for validation and 1000 images that can be used for testing. For the experiments using this dataset, we use the ASTRA toolbox [41–43] to simulate a parallel beam ray transform \mathcal{R} with 50 uniformly spaced views at angles between 0 and π . We simulate the measurements y as post-log data in a low-dose setting:

$$y = -\frac{1}{\mu} \log \left(\max \left\{ \frac{n}{N_{\text{in}}}, \eta \right\} \right), \quad \text{where } n \sim \text{Pois}(N_{\text{in}} \exp(-\mu \mathcal{R}(u))).$$

Here $N_{\text{in}} = 10\,000$ is the average number of photons per detector pixel (without attenuation), μ is a base attenuation coefficient connecting the volume geometry and attenuation strength, and η is a small constant to ensure that the argument of the logarithm is strictly positive, chosen as $\eta = 10^{-8}$ in our experiments. Figure 4 shows some examples of the ground truth images and filtered backprojection (FBP) reconstructions from the corresponding simulated measurements. In these experiments, we will define the data discrepancy functional E_y as

$$E_y(u) = \frac{1}{2} \|\mathcal{R}u - y\|_2^2.$$

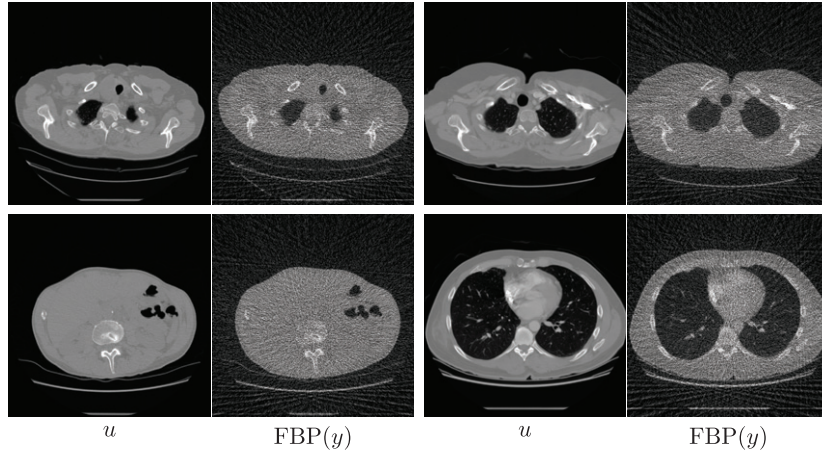


Figure 4. Four samples of the images that were used to train the reconstruction operators in the CT experiments, and the results of applying FBP to the corresponding simulated sinograms. The images are clipped between upper and lower attenuation coefficient limits of -1024 HU and 1023 HU.

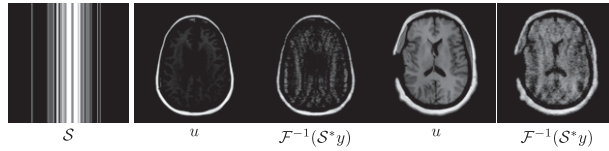


Figure 5. The sampling mask S used in the MRI experiments, sampling 20.3% of k -space, and two samples of the images that were used to train the reconstruction operators in the MRI experiments, and the zero-filling reconstructions from the corresponding simulated k -space measurements.

5.1.2. FastMRI. We use a selection of axial T1-weighted brain images of size 320×320 from the FastMRI dataset [44, 45] for our MRI experiments. As in section 5.1.1, we screen the images to remove as many low-quality images as possible. The set is split into 5000 images that can be used for training, 200 images that can be used for validation and 1000 images that can be used for testing. For the experiments using this dataset, we simulate the measurements using a discrete Fourier transform \mathcal{F} and a variable density Cartesian line sampling pattern S (simulated using the software package associated with the work in [46] and shown in figure 5):

$$y = \mathcal{S}\mathcal{F}u + \varepsilon,$$

where ε is complex-valued white Gaussian noise. In this setting, a complex-valued image is modelled as a real image with two channels, one for the real part and the other for the imaginary part. The corresponding data discrepancy functional (E_y in equation (9)) will be defined as

$$E_y(u) = \frac{1}{2} \|\mathcal{S}\mathcal{F}u - y\|_2^2.$$

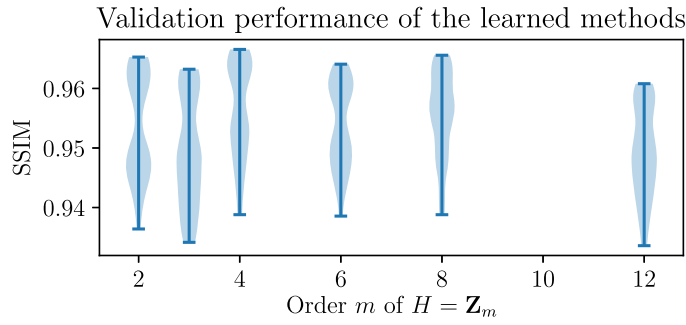


Figure 6. The reconstruction quality, as measured on a validation set, of learned proximal gradient methods trained on the CT reconstruction problem with varying orders of the group H . Note that when H is chosen to represent on-grid rotations (i.e. $m = 2$ or $m = 4$), the performance is significantly better than for any of the other choices of H .

5.2. Experimental setup

5.2.1. Learning framework. Although it is also possible to learn the parameters of the reconstruction methods in algorithm 2 in an unsupervised learning setting, all experiments that we consider in this work can be classified as supervised learning experiments: given a finite training set $\{(u_i, y_i)\}_{i=1}^N$ of ground truth images u_i and corresponding noisy measurements y_i , we choose the parameters of Φ in algorithm 2 by solving the empirical risk minimisation problem

$$\min_{\Phi} \frac{1}{N} \sum_{i=1}^N \|u_i - \Phi(y_i)\|_2^2.$$

5.2.2. Architectures and initialisations of the reconstruction networks. We use the reconstruction networks defined in section 4.2, referring to the architecture described there with roto-translationally equivariant components as the equivariant method and referring to the architecture with translationally equivariant components as the ordinary method. To ensure fair comparisons between the various methods that we compare, we fix as many as possible of the aspects of the methods that are orthogonal to the point investigated in the experiments. To this end, every learned proximal gradient method has a depth of $\text{it} = 8$ iterations. Both for the CT and MRI experiment, the images being recovered are two-dimensional, so we use equivariant convolutions with respect to groups of the form $\mathbf{R}^2 \rtimes \mathbf{Z}_m$. Since the equivariant convolutions are implemented using ordinary convolutions, it is natural and straightforward to compare methods with the same width. The width of each network is the same (feature vectors that transform according to the regular representation take up $|H|$ ‘ordinary’ channels, and we fix the size of the product $|H| \cdot n_{\text{channels}} = 96$ where n_{channels} is the number of such feature vectors in the intermediate part of $\widehat{\text{prox}}_i$ in equation (11)). All convolution filters used are of size 3×3 . We choose the initial reconstruction $u^0 = 0$ and use a memory variable s of five scalar channels wide in the learned proximal gradient method (algorithm 2).

Furthermore we ensure that the initialisation of both types of methods are comparable. Referring back to equation (11), we choose to initialise $K_{\text{intermediate},i}$ equal to zero and let $K_{\text{project},i}$ and $K_{\text{lift},i}$ be randomly initialised using the He initialisation method [47], as implemented in PyTorch [48] for ordinary convolutions and generalised to equivariant convolutions in [26] and implemented in the software package <https://github.com/QUVA-Lab/e2cnn> [11]. For the practical implementation of the exact methods studied, the reader is advised to consult the code at https://github.com/fsherry/equivariant_image_recon [54].

5.2.3. Hyperparameters of the equivariant methods. In addition to the usual parameters of a CNN, the learned equivariant reconstruction methods have additional parameters related to the choice of the symmetry group its representations to use. In this work, we have chosen to work with groups of the form $\mathbf{R}^2 \rtimes \mathbf{Z}_m$, so a choice needs to be made which $m \in \mathbf{N}$ to consider.

In figure 6, we see the result of training and validating learned equivariant reconstruction methods on the CT reconstruction problem, with various orders m of the group $H = \mathbf{Z}_m$. Each of the learned methods is trained on the same training set consisting of 100 images. The violin plots used give kernel density estimates of the distributions of the performance measures; for each one, we have omitted the top and bottom 5% of values so as not to be misled by outliers. Evidently, in this case, the groups of on-grid rotations significantly outperform the other choices, with $m = 4$ giving the best performance. Based on this result, all further experiments with the equivariant methods will use the group $H = \mathbf{Z}_4$.

5.2.4. Training details. For both the equivariant and ordinary reconstruction methods, we train the methods using the Adam optimisation algorithm [49] with learning rate 10^{-4} , $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 10^{-8}$. We use minibatches of size 1 and perform a total of 10^5 iterations of the Adam algorithm to train each method, so that we perform the same total number of iterations for each training set, regardless of its size. Since we have chosen to use the finite group approach, with intermediate fields transforming according to their regular representation, we can use a pointwise nonlinearity for both the equivariant and ordinary reconstruction methods. In all experiments, we use the leaky ReLU function as the nonlinearity (ϕ in equation (11)), applied pointwise:

$$\phi(x) = \begin{cases} x & \text{if } x > 0, \\ 0.01x & \text{else.} \end{cases}$$

Each training run is performed on a computer with an Intel Xeon Gold 6140 CPU and an NVIDIA Tesla P100 GPU. Training the equivariant methods requires slightly more computational effort than the ordinary methods: to begin with, given the specification of the architecture, bases need to be computed for the equivariant convolution kernels (this takes negligible effort compared to the effort expended in training). Besides this, each training iteration requires the computation of the convolutional filter from its parameters and the basis functions and the back-propagation through this basis expansion. To give an example of the extra computational effort required, we have timed 100 training iterations for comparable equivariant and ordinary methods for the MRI reconstruction problem: this took 35.5 s for the ordinary method and 41.9 s for the equivariant method, an increase of 18%. These times correspond to a total training time of 9.9 h and 11.6 h for the equivariant and ordinary methods respectively. Note that at test time, however, the ordinary and equivariant methods can be computed with the same effort.

5.2.5. Performance measures. We evaluate the performance of the learned reconstruction methods using two performance measures:

- The peak signal to noise ratio (PSNR), defined for a ground truth signal $u \in \mathbf{R}^n$ and reconstruction $\hat{u} \in \mathbf{R}^n$ as

$$\text{PSNR}(\hat{u}, u) = 10 \log_{10} \left(\frac{n \max_{1 \leq i \leq n} |u_i|^2}{\|u - \hat{u}\|^2} \right).$$

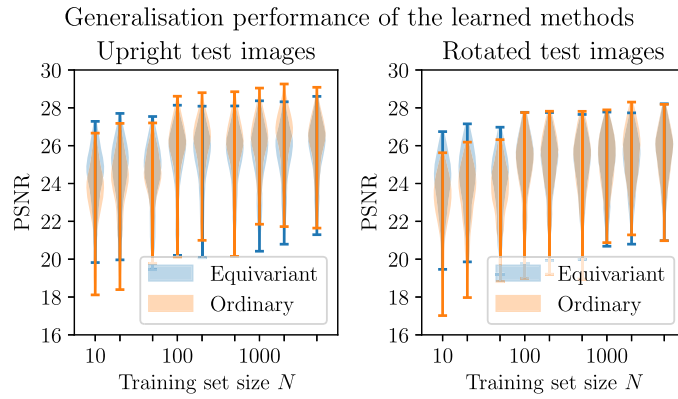


Figure 7. A comparison of the performance of equivariant and ordinary learned proximal gradient methods trained on training sets of various sizes for the CT reconstruction problem. The methods are tested on images that have not been seen during training time, both in the same orientations as were observed during training (‘upright test images’) and rotated at random angles (‘rotated test images’). The performance is evaluated on the lung regions only.

- The structural similarity index measure (SSIM) [50], defined initially on small windows of images, $u, \hat{u} \in \mathbf{R}^{w \times w}$ (with w odd) by

$$\text{SSIM}(\hat{u}, u) = \frac{2\bar{\hat{u}}\bar{u} + c_1}{\hat{u}^2 + \bar{u}^2 + c_1} \cdot \frac{2s_{\hat{u},u} + c_2}{s_{\hat{u}}^2 + s_u^2 + c_2}$$

for small nonnegative constants c_1, c_2 . In this formula, we have used the mean and variance statistics defined by

$$\bar{u} = \frac{1}{w^2} \sum_{1 \leq i,j \leq w} u_{i,j}, \quad s_{\hat{u},u} = \frac{1}{w^2} \sum_{1 \leq i,j \leq w} (\hat{u}_{i,j} - \bar{\hat{u}})(u_{i,j} - \bar{u}), \quad s_u = s_{u,u}.$$

To obtain a performance measure for larger images $u, \hat{u} \in \mathbf{R}^{n_1 \times n_2}$ with $n_1, n_2 \geq w$, we compute the SSIM on each of their subwindows and average:

$$\text{SSIM}(\hat{u}, u) = \frac{1}{(n_1 + 1 - w)(n_2 + 1 - w)} \sum_{\substack{1 \leq i \leq n_1 + 1 - w \\ 1 \leq j \leq n_2 + 1 - w}} \text{SSIM}([\hat{u}]_{i,j}, [u]_{i,j}),$$

where $[u]_{i,j}$ is the window $(u_{k,l})_{i \leq k < i+w, j \leq l < j+w}$. We use the implementation included in scikit-image [51], with the corresponding default parameter choices.

Both the PSNR and the SSIM have the property that higher values correspond to better reconstructions. Whereas the PSNR can immediately be applied to arbitrarily shaped signals (since the various locations in the signals do not interact), the SSIM in principle requires the input images to be regularly sampled to make sense of the subwindow statistics. One way in which the SSIM can be reasonably computed on segmented data is as follows: note that the subwindow SSIMs that are needed in the computation of the full SSIM define an image, the so-called SSIM map. If the input images are first padded on each side by $\lfloor w/2 \rfloor$ pixels (for example by reflection padding, as is done in the scikit-image implementation), the SSIM map computed from them will be of the same size as the original input images and will be aligned

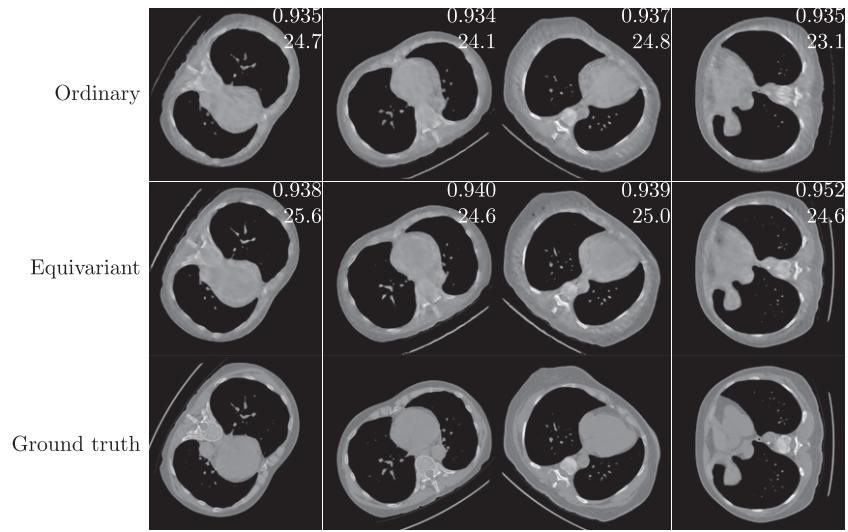


Figure 8. A random selection of test images corresponding to the plots shown in figure 7, with a training set of size $N = 50$. On each reconstruction, the top number is its SSIM and the bottom number is its PSNR w.r.t. the ground truth, with both performance measures computed on the lung regions only. The images are clipped between upper and lower attenuation coefficient limits of -1024 HU and 1023 HU.

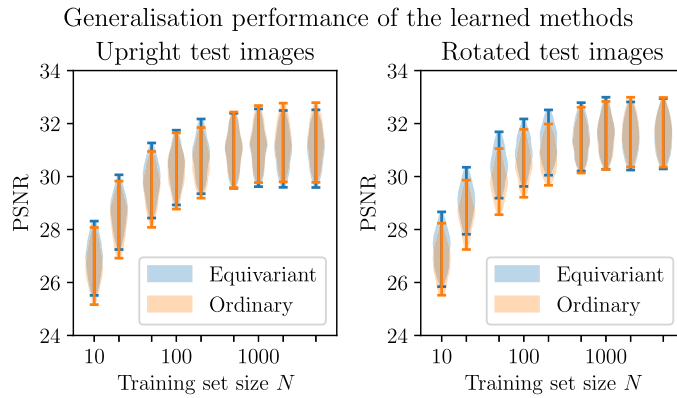


Figure 9. A comparison of the performance of equivariant and ordinary learned proximal gradient methods trained on training sets of various sizes for the MRI reconstruction problem. The methods are tested on images that have not been seen during training time and that have been rotated at random angles. The performance is evaluated on the foreground regions only.

with them. The ordinary SSIM is computed by taking the average of such an SSIM map, so given a segmentation mask we can compute a segmented SSIM by instead taking the average of the values of the SSIM map over points that are inside the mask.

To apply either of the performance measures to the MRI images, which are complex-valued, we compute them on the absolute value images.

5.3. CT experiment: varying the size of the training set

In this experiment, we study the effect of varying the size of the training set on the performance of the equivariant and ordinary methods. We consider a range of training set sizes, as shown in figure 7, and test the learned reconstruction methods on images that were not seen during training time, both in the same orientation and randomly rotated images. In medical applications, one tends to be particularly interested in the lung regions of the chest CT images. Although the methods have not been trained with this specifically in mind, in this section we will consider their performance on the lung regions. For this purpose, we use an automatic lung CT segmentation tool from [52] to select the regions of interest. As can be seen in figure 8, the equivariant method does a better job at reconstructing the lung regions than the ordinary method when trained on smaller training sets, but does slightly worse with larger training sets. This can be explained by the fact that the equivariant method is subsumed by the ordinary method (recall that the equivariant method can be replicated by appropriately setting the weights of the ordinary method, but the converse does not hold). The violin plots displayed have the same interpretation as those shown in figure 6 and described in section 5.2.3. We see a slight deviation from a monotonic relationship between the training set size and reconstruction quality that would usually be expected. Small random variations in the test performance can be explained by various nondeterministic aspects of the training procedure: we use random initialisations of the network weights, the learning problem is nonconvex and there is randomness in how the examples are sampled during training. From this comparison, we see that the equivariant method is able to better take advantage of smaller training sets than the ordinary method. Furthermore, we see that the performance of the equivariant method does not suffer much when testing on images in unseen orientations, whereas the performance of the ordinary method drops significantly when testing on rotated images. Figure 8 shows some examples of test reconstructions made with the methods learned on a training set of size $N = 50$. In these reconstructions, it can be seen that the equivariant method does better at removing streaking artefacts than the ordinary method.

5.4. MRI experiment: varying the size of the training set

This experiment is similar to the experiment in section 5.3, but concerns the MRI reconstruction problem. A notable difference with the CT reconstruction problem is that, as a result of the Cartesian line sampling pattern, the forward operator is now less compatible with the rotational symmetry. Regardless of this, we have seen in section 4 that it is still sensible in this context to use equivariant neural networks in a method motivated by a splitting optimisation method. As in section 5.3, we evaluate the performance of the learned methods on regions of interest: in this case we use the foreground of the images, which we isolate by thresholding the ground truth images, followed by taking the convex hull of the result. The performance differential between the equivariant and ordinary methods is more subtle than in the CT reconstruction problems. An explanation for this can be found in the fact that the MRI reconstruction problem is, in a certain sense, easier than the CT reconstruction problem: the nonzero singular values of the MRI forward operator are constant, while those of the CT forward operator decay, complicating the inversion. Remarkably, it is observed that both methods perform better on the rotated images than they do on the upright images. This is an artefact of how the rotated images are created: rotated images are generated from the upright images by performing a rotation operation which necessarily includes an interpolation step. As a result of this, some of the high frequency details disappear after rotating, resulting in an easier reconstruction problem. Appendix A goes into more detail about this effect. In figure 9, we see that the equivariant method can again take better advantage of smaller training sets and is more robust to images dissimilar to those seen

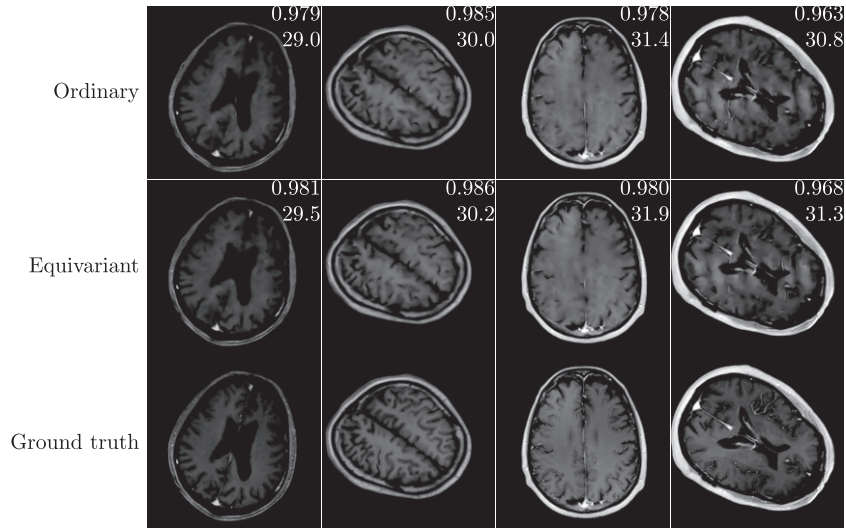


Figure 10. A random selection of test images corresponding to the plots shown in figure 9, with a training set of size $N = 100$. On each reconstruction, the top number is its SSIM and the bottom number is its PSNR w.r.t. the ground truth, with both performance measures computed on the foreground regions only.

in training. Figure 10 shows examples of reconstructions made with the methods learned on a training set of size $N = 100$.

6. Conclusions and discussion

In this work, we have shown that equivariant neural networks can be naturally incorporated into learnable reconstruction methods for inverse problems. Doing so requires little extra effort and results in higher quality reconstructions when compared to similar methods that use ordinary CNNs. The main difference of this approach compared to existing approaches is that we model proximal operators in a learned reconstruction method as roto-translationally equivariant rather than just translationally equivariant, as is usually the case. Building the extra symmetries into the learned reconstruction method has the effect of lowering the method's sample complexity. Using roto-translationally equivariant neural networks as opposed to ordinary CNNs results in better performance when trained on smaller training sets and more robustness to rotations.

Let us now discuss some of the limitations of the proposed approach and potential improvements to be considered in future work.

As we saw in sections 5.3 and 5.4, the equivariant method outperforms the ordinary method for small training sets, but is slightly outperformed by the ordinary method for large training sets. This is a result of the equivariant method being a subset of the ordinary method. The equivariant method can be made more expressive by using a larger number of intermediate channels, but this comes at the expense of increased computational cost.

In section 5.2.3, we saw that the learned methods perform best when the group H is chosen to be a group of on-grid rotations. In theory, one would expect better performance with a larger number of rotations, but in practice there is the issue of how the equivariant kernels are discretised. Indeed, when solving the constraint for equivariance in equation (8), the allowed kernels turn out to be circular harmonics multiplied by an arbitrary radial profile, and in practice we discretise these functions on 3×3 filters. An opportunity for future work

on the use of equivariant neural networks can be found in how the combination of group and discretisation should be optimised.

All of the experiments shown in this work have dealt with two-dimensional images, but the methods described here can be applied equally well to three-dimensional images, as long as the two-dimensional equivariant convolutions are replaced by their three-dimensional counterparts. The representation theory of $SO(3)$ is more complicated than that of $SO(2)$, but it is similarly possible to design roto-translationally equivariant convolutions in three dimensions [27]. One potential application is mentioned in remark 1: in diffusion tensor MRI, the domain is three-dimensional, with the additional challenge that the image that is to be recovered is a tensor field rather than a scalar field.

In the experiments that we demonstrated in this work, we focused on a single type of learned reconstruction operator, the learned proximal gradient method. In fact, the framework that we describe is not limited to this form of reconstruction algorithm. As an example of another type of learned reconstruction operator, consider the learned primal-dual method of [53]. A small corollary to proposition 2 is that, when J is invariant and the Fenchel conjugate J^* is well-defined, prox_{J^*} will be equivariant in the same way that prox_J is. As a result, assuming reasonable invariance properties of a data discrepancy term, a learned primal-dual method can be considered where both the primal and dual proximal operators are modelled as appropriate equivariant neural networks.

Acknowledgments

Data used in the preparation of this article were obtained from the NYU fastMRI Initiative database (fastmri.med.nyu.edu) [44, 45]. As such, NYU fastMRI investigators provided data but did not participate in analysis or writing of this report. A listing of NYU fastMRI investigators, subject to updates, can be found at fastmri.med.nyu.edu. The primary goal of fastMRI is to test whether machine learning can aid in the reconstruction of medical images. The authors acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health, and their critical role in the creation of the free publicly available LIDC/IDRI Database used in this study [39, 40]. MJE acknowledges support from the EPSRC Grants EP/S026045/1 and EP/T026693/1, the Faraday Institution via EP/T007745/1, and the Leverhulme Trust fellowship ECF-2019-478. CE and CBS acknowledge support from the Wellcome Innovator Award RG98755. CBS acknowledges support from the Leverhulme Trust project on ‘Breaking the non-convexity barrier’, the Philip Leverhulme Prize, the EPSRC Grants EP/S026045/1 and EP/T003553/1, the EPSRC Centre No. EP/N014588/1, European Union Horizon 2020 research and innovation programmes under the Marie Skłodowska-Curie Grant agreement No. 777826 NoMADS and No. 691070 CHiPS, the Cantab Capital Institute for the Mathematics of Information and the Alan Turing Institute. FS acknowledges support from the Cantab Capital Institute for the Mathematics of Information. EC and BO thank the SPIRIT project (No. 231632) under the Research Council of Norway FRIPRO funding scheme.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: [10.5281/zenodo.4556410](https://doi.org/10.5281/zenodo.4556410).

Appendix A. The blurring effect of image rotations

In section 5.4, we made the remarkable observation that the learned reconstruction methods perform better for the MRI problem on rotated images than on upright images similar to those

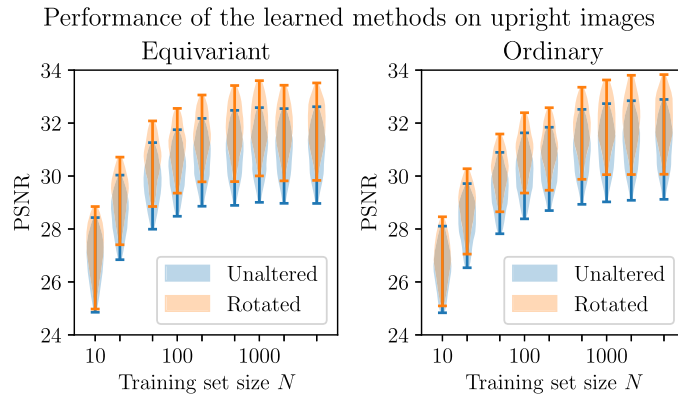


Figure A1. A comparison of the performance of the learned reconstruction methods on two types of upright images for the MRI problem: the original images ('unaltered') and otherwise identical images that have been rotated and rotated back ('rotated').

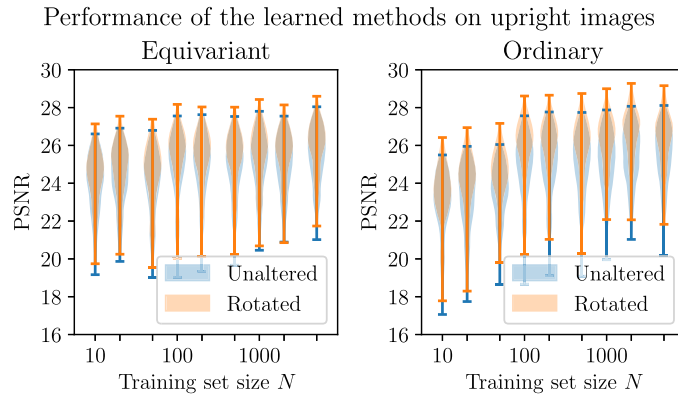


Figure A2. A comparison of the performance of the learned reconstruction methods on two types of upright images for the CT problem: the original images and otherwise identical images that have been rotated and rotated back.

on which they were trained. It was mentioned there that this is an artefact of the way in which rotated images are created. As a simple test of this explanation, consider the comparison of the performance on the unaltered upright images and the performance on upright images that have been randomly rotated and then rotated back to be upright. If the hypothesised explanation for the difference in performance is correct, we would expect the methods to perform better on the images that have been rotated and rotated back than on the unaltered images. Figure A1 shows the result of doing this comparison, confirming that the MRI problem is significantly easier to solve for the learned reconstruction methods after the images have undergone the blurring effect of the rotation operation. Figure A2 shows the same comparison repeated for the CT problem. In this case the effect is still visible, but it is considerably weaker, which explains why it was not observed in section 5.3.

ORCID iDs

Carola-Bibiane Schönlieb  <https://orcid.org/0000-0003-0099-6306>

Ferdia Sherry  <https://orcid.org/0000-0003-4809-7254>

References

- [1] Silver D *et al* 2016 Mastering the game of Go with deep neural networks and tree search *Nature* **529** 484–9
- [2] Senior A W *et al* 2020 Improved protein structure prediction using potentials from deep learning *Nature* **577** 706–10
- [3] Brown T B *et al* 2020 Language models are few-shot learners (arXiv:2005.14165[cs.CL])
- [4] Jin K H, McCann M T, Froustey E and Unser M 2017 Deep convolutional neural network for inverse problems in imaging *IEEE Trans. Image Process.* **26** 4509–22
- [5] Adler J and Öktem O 2017 Solving ill-posed inverse problems using iterative deep neural networks *Inverse Problems* **33** 124007
- [6] Lunz S, Öktem O and Schönlieb C-B 2018 Adversarial regularizers in inverse problems *Adv. Neural Inf. Process. Syst.* **31** 8516–25
- [7] Zandehshahvar M, Kiarashi Y, Zhu M, Maleki H, Brown T and Adibi A 2021 Manifold learning for knowledge discovery and intelligent inverse design of photonic nanostructures: breaking the geometric complexity (arXiv:2102.04454[physics.optics])
- [8] LeCun Y and Bengio Y 1998 Convolutional networks for images, speech, and time series *The Handbook of Brain Theory and Neural Networks* (Cambridge, MA: MIT Press)
- [9] Bekkers E J, Lafarge M W, Veta M, Eppenhof K A J, Pluim J P W, Duits R and Duits R 2018 Roto-translation covariant convolutional networks for medical image analysis *Medical Image Computing and Computer Assisted Intervention (MICCAI)* vol 2018 pp 440–8
- [10] Worrall D E, Garbin S J, Turmukhambetov D and Brostow G J 2017 Harmonic networks: deep translation and rotation equivariance *Proc. IEEE Conf. Computer Vision and Pattern Recognition* pp 5028–37
- [11] Weiler M and Cesa G 2019 General $E(2)$ -equivariant steerable CNNs *Advances in Neural Information Processing Systems* ed H Wallach, H Larochelle, A Beygelzimer, F d' Alché-Buc, E Fox and R Garnett vol 32 pp 14334–45
- [12] Cohen T S and Welling M 2016 Group equivariant convolutional networks *Proc. 33rd Int. Conf. Machine Learning* pp 2990–9
- [13] Dieleman S, De Fauw J and Kavukcuoglu K 2016 Exploiting cyclic symmetry in convolutional neural networks *Proc. 33rd Int. Conf. Machine Learning* pp 1889–98
- [14] Sosnovik I, Szmaja M and Smeulders A 2019 Scale-equivariant steerable networks (arXiv:1910.11093[cs.CV])
- [15] Worrall D E and Welling M 2019 Deep scale-spaces: equivariance over scale (arXiv:1905.11697[cs.LG])
- [16] Finzi M, Stanton S, Izmailov P and Wilson A G 2020 Generalizing convolutional neural networks for equivariance to Lie groups on arbitrary continuous data (arXiv:2002.12880[stat.ML])
- [17] Hounsfield G N 1973 Computerized transverse axial scanning (tomography): Part 1. Description of system *Br. J. Radiol.* **46** 1026–2
- [18] Lauterbur P C 1973 Image formation by induced local interactions: examples employing nuclear magnetic resonance *Nature* **242** 190–1
- [19] Mansfield P and Grannell P K 1975 'Diffraction' and microscopy in solids and liquids by NMR *Phys. Rev. B* **12** 3618–34
- [20] Hadamard J 1902 Sur les problèmes aux dérivées partielles et leur signification physique *Princeton Univ. Bull.* **13** 49–52
- [21] Krizhevsky A, Sutskever I and Hinton G E 2012 ImageNet classification with deep convolutional neural networks *Advances in Neural Information Processing Systems* ed F Pereira, C J C Burges, L Bottou and K Q Weinberger vol 25 pp 1097–105
- [22] Putzky P and Welling M 2017 Recurrent inference machines for solving inverse problems (arXiv:1706.04008[cs.NE])

- [23] Engl H W, Hanke M and Neubauer A 1996 *Regularization of Inverse Problems (Mathematics and its Applications vol 375)* (Dordrecht: Kluwer)
- [24] Hansen P C 2010 *Discrete Inverse Problems (Insight and Algorithms)* (Philadelphia, PA: SIAM)
- [25] Cohen T S, Geiger M and Weiler M 2019 A general theory of equivariant CNNs on homogeneous spaces *Advances in Neural Information Processing Systems* vol 32 pp 9145–56
- [26] Weiler M, Hamprecht F A and Storath M 2018 Learning steerable filters for rotation equivariant CNNs *Proc. IEEE Conf. Computer Vision and Pattern Recognition* pp 849–58
- [27] Weiler M, Geiger M, Welling M, Boomsma W and Cohen Taco S 2018 3D steerable CNNs: learning rotationally equivariant features in volumetric data *Advances in Neural Information Processing Systems* vol 32 (Red Hook, NY: Curran Associates) pp 10381–92
- [28] Folland G B 2015 *A Course in Abstract Harmonic Analysis* 2nd edn (Boca Raton, FL: CRC Press)
- [29] Burger M and Osher S 2004 Convergence rates of convex variational regularization *Inverse Problems* **20** 1411–21
- [30] Bruck R E 1977 On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space *J. Math. Anal. Appl.* **61** 159–64
- [31] Passty G B 1979 Ergodic convergence to a zero of the sum of monotone operators in Hilbert space *J. Math. Anal. Appl.* **72** 383–90
- [32] Beck A and Teboulle M 2009 A fast iterative shrinkage-thresholding algorithm for linear inverse problems *SIAM J. Imag. Sci.* **2** 183–202
- [33] Moreau J J 1962 Fonctions convexes duales et points proximaux dans un espace hilbertien *Comptes Rendus Hebdomadaires Des Séances De L'Académie Des Sciences* vol 255 (Paris: Gauthier-Villars) pp 2897–9
- [34] Moreau J J 1963 Propriétés des applications ‘prox’ *Comptes Rendus Hebdomadaires Des Séances De L'Académie Des Sciences* vol 256 (Paris: Gauthier-Villars) pp 1069–71
- [35] Moreau J J 1965 Proximité et dualité dans un espace hilbertien *Bul. Soc. Math. France* **79** 273–99
- [36] Rudin L I, Stanley O and Fatemi E 1992 Nonlinear total variation based noise removal algorithms *Physica D* **60** 259–68
- [37] Bredies K, Kunisch K and Pock T 2010 Total generalized variation *SIAM J. Imag. Sci.* **3** 492–526
- [38] Coulon O, Alexander D C and Arridge S 2004 Diffusion tensor magnetic resonance image regularization *Med. Image Anal.* **8** 47–67
- [39] Armato S G III et al 2011 The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans *Med. Phys.* **38** 915–31
- [40] Armato S G III et al 2015 Data from LIDC-IDRI [Data set] *The Cancer Imaging Archive* <https://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX>
- [41] Palenstijn W J, Batenburg K J and Sijbers J 2011 Performance improvements for iterative electron tomography reconstruction using graphics processing units (GPUs) *J. Struct. Biol.* **176** 250–3
- [42] van Aarle W, Palenstijn W J, De Beenhouwer J, Altantzis T, Bals S, Joost Batenburg K and Sijbers J 2015 The ASTRA toolbox: a platform for advanced algorithm development in electron tomography *Ultramicroscopy* **157** 35–47
- [43] van Aarle W, Palenstijn W J, Cant J, Janssens E, Bleichrodt F, Dabrovolski A, De Beenhouwer J, Joost Batenburg K and Sijbers J 2016 Fast and flexible x-ray tomography using the ASTRA toolbox *Opt. Express* **24** 25129–47
- [44] Knoll F et al 2020 fastMRI: a publicly available raw k -space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning *Radiology* **2** e190007
- [45] Zbontar J et al 2019 fastMRI: an open dataset and benchmarks for accelerated MRI (arXiv:1811.08839[cs.CV])
- [46] Lustig M, Donoho D and Pauly J M 2007 Sparse MRI: the application of compressed sensing for rapid MR imaging *Magn. Reson. Med.* **58** 1182–95
- [47] He K, Zhang X, Ren S and Sun J 2015 Delving deep into rectifiers: surpassing human-level performance on ImageNet classification *Proc. IEEE Int. Conf. Computer Vision* pp 1026–34
- [48] Paszke A et al 2019 PyTorch: an imperative style, high-performance deep learning library *Advances in Neural Information Processing Systems* vol 32 (Red Hook, NY: Curran Associates) pp 8026–37
- [49] Kingma D P and Ba J 2017 Adam: a method for stochastic optimization (arXiv:1412.6980)
- [50] Brunet D, Vrsay E R and Wang Z 2012 On the mathematical properties of the structural similarity index *IEEE Trans. Image Process.* **21** 1488–99

- [51] van der Walt S, Schönberger J L, Nunez-Iglesias J, Boulogne F, Warner J D, Yager N, Gouillart E and Yu T 2014 scikit-image: image processing in Python *Peer J.* **2** e453
- [52] Hofmanninger J, Prayer F, Pan J, Röhrich S, Prosch H and Langs G 2020 Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem *Eur. Radiol. Exp.* **4** 50
- [53] Adler J and Öktem O 2018 Learned primal-dual reconstruction *IEEE Trans. Med. Imag.* **37** 1322–32
- [54] Ferdia S 2021 fsherry/equivariant_image_recon v0.1.0, Zenodo, <https://doi.org/10.5281/zenodo.4556410>